Machine Learning
Homework Assignment 3

Nishika Chopra Karanpreet Singh Wadhwa	nc2259 N18598308 ksw352 N13337853

ans 1) (a) Example 1:
4 +ve / 5 -ve

Example 2:
3 +ve / 6 -ve

High entropy means that our data labels are equally ~~distributed~~ ~~and~~ ~~categorized~~ / present

in the examples, Here,
in Example 1 we see that no. of positive and negatives is almost equal as compared to example 2
Hence entropy of Example 1 is higher than entropy of Example 2.

This can be validated using the formula

$$S_1 = -\frac{4}{9}\log\left(\frac{4}{9}\right) - \frac{5}{9}\log\left(\frac{5}{9}\right)$$

$$= \cancel{0.2979}\ 0.992$$

$$S_2 = -\frac{3}{9}\log\left(\frac{3}{9}\right) - \frac{6}{9}\log\left(\frac{6}{9}\right)$$

$$= \cancel{0.917}\ \cancel{0.277}\ 0.92$$

$$S_1 > S_2$$

Hence Proved

(b) Data Set

	x_1	x_2	x_3	γ
x^1	F	F	F	+
x^2	T	F	T	+
x^3	F	F	T	−
x^4	F	T	F	+
x^5	T	T	F	−
x^6	T	T	T	−
x^7	F	F	F	−

we split on x_1

	x_1	x_2	x_3	γ
x^1	F	F	F	+
x^3	F	F	T	−
x^4	F	T	F	+
x^7	F	F	F	−

	x_1	x_2	x_3	γ
x^2	T	F	T	+
x^5	T	T	F	−
x^6	T	T	T	−

$P_+(S) = \frac{3}{7}$ $\quad P_-(S) = \frac{4}{7}$

0.918

Entropy (S) $= -\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right)$

$= 0.9863$

Scanned by CamScanner

$$P_-(x_{1F}) = \frac{2}{4} = \frac{1}{2} \qquad P_+(x_{1F}) = \frac{2}{4} = \frac{1}{2}$$

$$P_-(x_{1T}) = \frac{2}{3} \qquad P_+(x_{1T}) = \frac{1}{3}$$

<u>Entropy</u>

$$E(x_{1F}) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)$$

$$= 1$$

$$E(x_{1T}) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right)$$

$$= 0.924$$

Hence,

Information Gain

$$I\text{-}G(S) = -\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right) \cancel{-\frac{4}{7}}$$

$$- \left[\frac{4}{7}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) + \frac{3}{7}\left(-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right)\right]$$

$$= 0.9863 - 1 - 0.924$$

$$= 0.0203$$

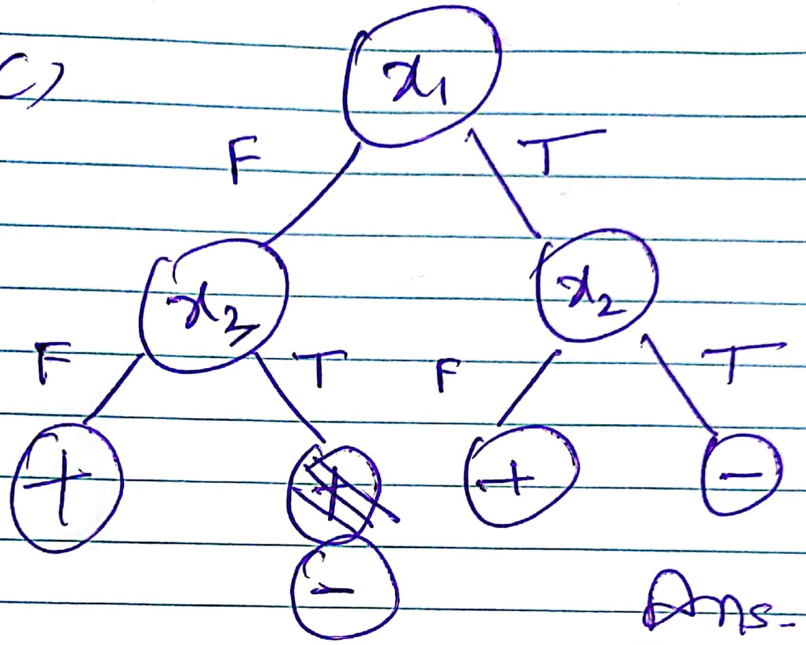
<u>Ans.</u>

Scanned by CamScanner

~~6.017~~

~~0.1426~~

(C)



~~Ans.~~

- First we check the entropies on each Attribute after split & Make a split on x_1 & it gives the least entropy After split.
- For Left side of x_1, x_3 gives the lesser entropy (i.e 0 entropy). So we Make split on x_3
- For Right side of x_1, x_2 gives the lesser entropy (i.e. 0 entropy). Thus we make a split on x_2.

(0% training error is Achieved)

Note: we have used the revised dataset with no training error

(d) $H(Y) = -P(+)\log[P(+)] - P(-)\log[P(-)]$

$= -\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right)$

$= 0.9863$

$H(Y|X) = \sum_{x} P[X=x] \times \left[\sum_{y} -P[Y=y \mid X=x] \times \log P[Y=y \mid X=x]\right]$

$= \frac{4}{7}\left[\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right)\right] + \frac{3}{7}\left[-\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)\right]$

$= 0.571 + 0.396$
$= 0.966$

$H(Y) - H(Y|X) = 0.9863 - 0.966$
$= 0.0203$ Ans.

(e) $E(s) = -\sum \frac{N_\ell}{N}\log\left(\frac{N_\ell}{N}\right)$

$= -\sum_{l \in L} p(l)\log p(l)$

Here, $P(l) = \frac{1}{z}$ since we have z possible labels

$E(s) = -\sum^{z} \frac{1}{z}\log\frac{1}{z} = \boxed{\log(z)}$

$= z\left[\frac{1}{z}\log(z)^{-1}\right]$

Ans.